SELECTING TV NEWS STORIES AND NEWSWIRE ARTICLES RELATED TO A TARGET ARTICLE OF NEWSWIRE USING SVM

Yoshimi Suzuki, Fumiyo Fukumoto and Yoshihiro Sekiguchi

Department of Computer Science and Media Engineering
Yamanashi University
4-3-11 Takeda, Kofu 400-8511 Japan

ABSTRACT

This paper describes a method for selecting TV news stories and newswire articles related to a target article of newswire by using a machine learning technique called SVM (Support Vector Machines). We used selected antecedents of overt pronouns, compound nouns in the experiments. The results of experiments showed that the use of antecedents of overt pronouns and compound nouns for SVM is effective. And SVM is more effective than term weighting methods such as word density, $TF * IDF$, χ^2, a method based on entropy or a method based on standard deviation for event detection.

1. INTRODUCTION

An item of news is reported by various news media, such as newspaper, newswire, news site on the web, TV news and radio news. It is useful to select various news media related to an event in order to retrieve information and summarize multi-documents.

In this paper, we propose a method for selecting newswire and TV news related to a target article of newswire (TAN). Figure 1 shows our system.

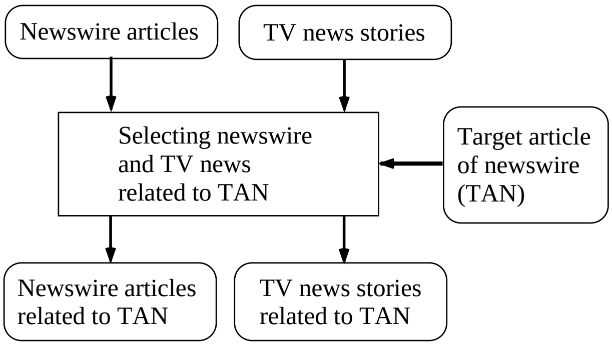
Newswire articles
TV news stories
Selecting newswire and TV news related to TAN
Target article of newswire (TAN)
Newswire articles related to TAN
TV news stories related to TAN

Figure 1: The system which selects TV news and newswire related to a target article of newswire (TAN)

However, newswire article and TV news story have some different features shown as follows:

Newswire article:

- Each newswire article is short.
- There are few keywords.
- Facts of each event are reported.

TV news:

- There are lots of synonyms which show a keyword.
- There are lots of words which aren't keywords.
- There are lots of overt pronouns.
- Various commentators or correspondents comment on the event.

Table 1 shows the compound word "World Trade Center" and the event "WTC Bombing trial"(WTC indicates "World Trade Center"). The compound word "World Trade Center" did not appear in 10 stories out of 18 stories about "WTC Bombing trial". Also the compound word "World Trade Center" appeared in some stories whose events are not "WTC Bombing trial". Therefore, it is more difficult to select the CNN news stories related to the target article of newswire than the Reuters newswire articles related to the target article of newswire.

Table 1: The compound word "World Trade Center" and the event "WTC Bombing trial"

Media	A/B	A/C
Reuters newswire	6/6=1.000	6/16=0.375
CNN TV news	8/18=0.444	8/60=0.133

A : the number of stories in which the compound word "World Trade Center" appear in B.
B : the number of stories about the event "WTC Bombing trial".
C : the number of stories in which the compound word "World Trade Center" appear.

Table 1 shows that it is difficult to select TV news related to the target newswire using small training data. Therefore, firstly we select newswire articles related to the target newswire article, then we select TV news stories related to TAN using the results of selecting newswire articles.

2. RELATED WORK

Our work is based on topic detection. There are many studies on topic detection [1]. However, most studies didn't mentioned overt pronoun resolver. In our work, we used overt pronoun resolver for selecting TV news related to a target article of newswire, and we showed the effectiveness of our overt pronoun resolver. Liddy studied effect of anaphora on retrieval results using WDF and IDF as term-weighting schemes and in her experiment the anaphora resolution was not effective [2]. We reported effect of overt pronoun resolution using SVM on event detection.

There are many studies on overt pronoun resolution. However, most studies are not simple, since they used semantic information. Our method is very simple and obtained comparative results.

In our prior study [3], we used some term weighting methods, i.e. word frequency, word density, $TF * IDF$, χ^2, entropy and a method based on standard deviation. We compared those term weighting methods and our method using SVM in Section 6.

Taira studied feature selection in SVM text categorization [4]. He suggested full number of words as features of SVM. We added compound nouns for features of SVM in the experiments.

3. AN OVERVIEW OF OUR SYSTEM

Figure 2 illustrates an overview of our system.

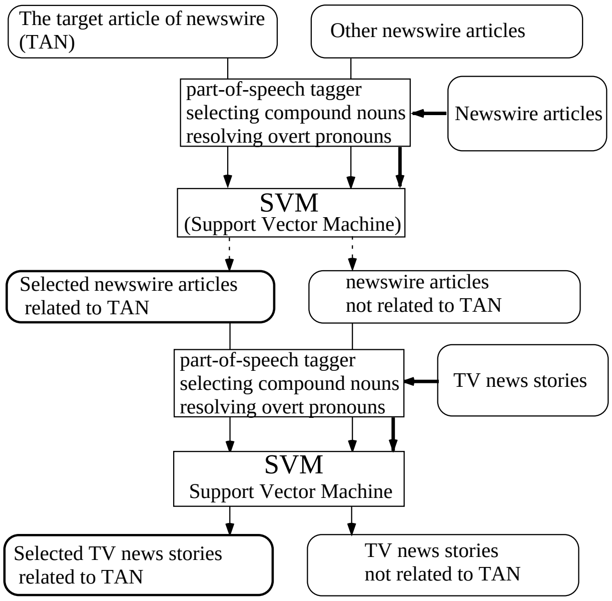

Figure 2: An Overview of Our System

In our system, firstly, we select newswire articles related to a target article and using the selected newswire articles we select TV news stories related to the target article using SVM (Support Vector Machine).

3.1. Preprocess of SVM

For preprocess of SVM, we performed tagging part-of-speech [5], extracting compound nouns and resolving overt pronouns. Compound nouns were extracted by some rules and the antecedents of overt pronoun selected by a noun-pronoun corresponding table [3].

3.2. Selecting newswire articles related to TAN

We used SVM for selecting newswire articles related to the target article of newswire. As positive example, we used the target article of newswire. As negative example, we used 300 of newswire articles whose events are not in the TDT 25 events.

3.3. Selecting TV news related to TAN

In TV news, there are lots of overt pronouns (25 of each news story), and the antecedents of most of them are keywords. For resolving overt pronouns, we used overt pronoun resolver which use noun-pronoun correspondence table. The table is made automatically using Brill's part of speech tagger [5] and the Reuters newswire articles. As positive example, we used the newswire articles related to the target article of newswire, as negative example, we used newswire articles whose events are not the event.

4. APPLYING SVM TO OUR SYSTEM

Support Vector Machines (SVM) is a relatively new learning approach introduced by [6] for solving two-class pattern recognition problems. It is based on the Structural Risk Minimization principle for which error-bound analysis has been theoretically motivated. The method is defined over a vector space where the problem is to find a decision surface that "best" separates the data points in two classes. In order to define "best" separation, we need to introduce the "margin" between two classes, i.e. the distance between two classes. The SVM problem is to find the decision surface that maximizes the margin between the data points in a training set.

We applied SVM to selecting newswire articles or TV news stories related to the target article of newswire. The first step to apply SVM is to make training and test data from the corpus. We represent every article in the training data as a vector. Let x_i be a article. The vector representation of x_i is:

$$\vec{x_i} = (w_1, \cdots, w_n)$$

where n is the total number of words in the training and test data. We used 3 sets of the value of w_j $(1 \leq j \leq n)$. If the event of an article x_i is same as the target event, then the label of x_i is +1, and -1 for being not a target event. In a similar way, each test data is represented as a vector. We used linear function as kernel function.

5. EXPERIMENTS

5.1. Data for Experiments

We used the corpus of TDT1 [7] in the experiments. There are 7,965 of articles of the Reuters and 7,898 of stories of the CNN TV news which were transcribed into texts in the corpus of TDT1. We carried out some experiments in order to make sure the effectiveness of our method using these data. For selecting newswire articles using SVM, we used 1 newswire article as positive example and 300 newswire articles as negative examples. For selecting TV news using SVM, we used the newswire articles whose event is same as the event of TAN as positive example and we used the newswire articles whose events are not same as the event of TAN as negative examples.

We used 5 sets of feature for selecting newswire and TV news shown as follows.

A : all words
B : all nouns
C : all nouns or compound nouns
D : all nouns and compound nouns
E : all nouns, compound nouns, antecedent of overt pronouns and all verbs (proposed method)

We used 3 sets of element value w_j of vectors which we used are shown as follows.

b : $w_j = 1$ if $word_j$ appears in the story.
$w_j = 0$ otherwise
f : $w_j =$ the number of $word_j$ in the story.
d : $w_j =$ density of $word_j$ in the story.

We used 4 different cut off numbers, i.e. 1, 3, 5, 10.

5.2. The results of selecting newswire articles related to the TAN

Table 2 shows the results of SVM for selecting the Reuters newswire articles related to the target article of newswire. In Table 2, c/o, f/v, e/r, pre, rec, F1 illustrates cut off, feature value, error rate, precision, recall and F-measure, respectively. When we selected **D** (all nouns and compound nouns) as the feature, **f** (the number of $word_j$ in the story) as feature value and **3** as cut off number, the result was better than the results of other combinations. However, the results were not sufficient to use as the training data of the SVM which selects TV news stories related to TAN.

5.3. The results of selecting TV news stories related to the TAN

Table 3 shows the results of SVM for selecting CNN TV news stories related to the target article of newswire. When we selected **E** (all nouns, compound nouns and antecedent of overt pronouns) as the feature, **f** (the number of $word_j$ in the story) as feature value and **5** as cut off number, the result was better than the results of other combinations.

6. DISCUSSION

For selecting the newswire articles related to TAN, using compound nouns was effective. However, effect of resolving overt pronoun was not sufficient. In order to improve it, we have to select some newswire articles related to TAN as preprocess of SVM.

Table 2: The results (the Reuters newswire articles)

feature	c/o	f/v	e/r	pre	rec	F1
A	10	b	0.037	1.000	0.083	0.153
	10	f	0.036	1.000	0.090	0.165
	10	d	0.036	1.000	0.090	0.165
	5	b	0.037	1.000	0.083	0.153
	5	f	0.036	1.000	0.090	0.165
	5	d	0.036	1.000	0.090	0.165
	3	b	0.037	1.000	0.083	0.153
	3	f	0.036	1.000	0.090	0.165
	3	d	0.036	1.000	0.090	0.165
	1	b	0.037	1.000	0.083	0.153
	1	f	0.037	1.000	0.086	0.159
	1	d	0.036	1.000	0.090	0.165
B	10	b	0.036	1.000	0.090	0.165
	10	f	0.031	0.878	0.248	0.387
	10	d	0.036	0.938	0.103	0.186
	5	b	0.036	1.000	0.090	0.165
	5	f	0.032	0.877	0.245	0.383
	5	d	0.036	0.939	0.107	0.192
	3	b	0.036	1.000	0.090	0.165
	3	f	0.032	0.877	0.245	0.383
	3	d	0.036	0.939	0.107	0.192
	1	b	0.036	1.000	0.093	0.170
	1	f	0.031	0.887	0.245	0.384
	1	d	0.036	0.939	0.107	0.192
C	10	b	0.041	0.455	0.086	0.145
	10	f	0.040	0.500	0.138	0.216
	10	d	0.037	1.000	0.083	0.153
	5	b	0.041	0.431	0.086	0.144
	5	f	0.040	0.506	0.141	0.221
	5	d	0.037	1.000	0.083	0.153
	3	b	0.041	0.417	0.086	0.143
	3	f	0.040	0.482	0.138	0.214
	3	d	0.037	1.000	0.083	0.153
	1	b	0.037	1.000	0.083	0.153
	1	f	0.035	1.000	0.128	0.226
	1	d	0.037	1.000	0.086	0.159
D	10	b	0.036	1.000	0.090	0.165
	10	f	0.031	0.890	0.252	0.392
	10	d	0.036	0.938	0.103	0.186
	5	b	0.036	1.000	0.093	0.170
	5	f	0.031	0.901	0.252	0.394
	5	d	0.036	0.938	0.103	0.186
	3	b	0.036	1.000	0.093	0.170
	3	**f**	**0.031**	**0.902**	**0.255**	**0.398**
	3	d	0.036	0.938	0.103	0.186
	1	b	0.036	1.000	0.097	0.176
	1	f	0.032	0.905	0.231	0.368
	1	d	0.036	0.938	0.103	0.186
E	10	b	0.036	1.000	0.090	0.165
	10	f	0.033	0.919	0.197	0.324
	10	d	0.036	0.966	0.097	0.176
	5	b	0.037	1.000	0.086	0.159
	5	f	0.033	0.933	0.193	0.320
	5	d	0.036	0.966	0.097	0.176
	3	b	0.037	1.000	0.086	0.159
	3	f	0.033	0.933	0.193	0.320
	3	d	0.036	0.966	0.097	0.176
	1	b	0.037	1.000	0.083	0.153
	1	f	0.033	0.933	0.193	0.320
	1	d	0.036	0.966	0.097	0.176

Table 3: The results (CNN TV news stories)

feature	c/o	f/v	e/r	pre	rec	F1
A	10	b	0.025	0.670	0.673	0.671
	10	f	0.022	0.716	0.739	0.727
	10	d	0.023	0.987	0.431	0.600
	5	b	0.025	0.672	0.673	0.672
	5	f	0.022	0.719	0.738	0.728
	5	d	0.023	0.987	0.424	0.593
	3	b	0.025	0.673	0.675	0.674
	3	f	0.022	0.720	0.738	0.729
	3	d	0.023	0.987	0.424	0.593
	1	b	0.025	0.671	0.674	0.672
	1	f	0.023	0.712	0.733	0.722
	1	d	0.023	0.987	0.423	0.592
B	10	b	0.016	0.842	0.822	0.832
	10	f	0.009	0.882	0.887	0.884
	10	d	0.015	0.984	0.643	0.778
	5	b	0.016	0.842	0.843	0.842
	5	f	0.009	0.882	0.885	0.884
	5	d	0.015	0.984	0.640	0.775
	3	b	0.016	0.842	0.844	0.843
	3	f	0.009	0.883	0.886	0.885
	3	d	0.015	0.984	0.636	0.773
	1	b	0.016	0.844	0.840	0.842
	1	f	0.009	0.884	0.884	0.884
	1	d	0.015	0.984	0.644	0.779
C	10	b	0.027	0.701	0.751	0.725
	10	f	0.018	0.749	0.816	0.781
	10	d	0.022	0.990	0.462	0.630
	5	b	0.027	0.710	0.753	0.731
	5	f	0.018	0.755	0.820	0.786
	5	d	0.022	0.988	0.453	0.621
	3	b	0.027	0.752	0.753	0.752
	3	f	0.015	0.815	0.820	0.818
	3	d	0.021	0.987	0.473	0.640
	1	b	0.027	0.782	0.761	0.771
	1	f	0.014	0.841	0.815	0.828
	1	d	0.021	0.985	0.474	0.640
D	10	b	0.017	0.821	0.833	0.827
	10	f	0.009	0.879	0.898	0.888
	10	d	0.014	0.989	0.658	0.790
	5	b	0.017	0.832	0.831	0.831
	5	f	0.009	0.890	0.897	0.894
	5	d	0.014	0.990	0.653	0.787
	3	b	0.017	0.830	0.834	0.832
	3	f	0.009	0.889	0.899	0.894
	3	d	0.014	0.990	0.651	0.785
	1	b	0.017	0.829	0.839	0.834
	1	f	0.008	0.897	0.901	0.899
	1	d	0.014	0.994	0.653	0.789
E	10	b	0.014	0.852	0.874	0.863
	10	f	0.008	0.894	0.916	0.905
	10	d	0.012	0.996	0.713	0.831
	5	b	0.014	0.859	0.872	0.865
	5	f	0.008	0.894	0.916	0.905
	5	d	0.012	0.996	0.711	0.830
	3	b	0.014	0.857	0.875	0.866
	3	f	0.008	0.896	0.917	0.906
	3	d	0.012	0.996	0.707	0.827
	1	b	0.014	0.856	0.881	0.868
	1	**f**	**0.008**	**0.897**	**0.917**	**0.907**
	1	d	0.012	0.995	0.709	0.828

For selecting the TV news stories related to TAN, we earned sufficient result using compound nouns and antecedents of overt pronouns. In our prior study, we used term weighting (a method based on standard deviation). Table 4 shows the results using a method based on term weighting [3] and the results of our method using SVM. The result of our method is better than the result using a method based on term weighting. Because, we can use low frequency of words as features without over-fitting to data using SVM.

Table 4: Comparing our method with a method based on term weighting

Method	precision	recall	F-measure
Term weighting (standard deviation)	0.665	0.699	0.682
SVM	0.897	0.917	0.907

7. CONCLUSIONS

We developed a system which selects the newswire articles and TV news stories related to a target article of newswire using SVM whose features include resolved overt pronouns and compound nouns. Our system is useful for multi-document summarization or information retrieval of plural media.

8. ACKNOWLEDGMENTS

This work was supported by the Grants from the Government subsidy for aiding scientific researchers (No.11780257) of the Ministry of Education of Japan.

9. REFERENCES

1. Frederick Walls and Hubert Jin and Sreenivasa Sista and Richard Schwartz, "Topic Detection in Broadcast News", DARPA Broadcast News Workshop, http://www.itl.nist.gov/div894/894.01/proc/darpa99/, 1999
2. Elizabeth DuRoss Liddy "ANAPHORA IN NATURAL LANGUAGE PROCESSING AND INFORMATION RETRIEVAL", Information Processing & Management Vol.26, No.1, pp.93-52, 1990.
3. Yoshimi Suzuki and Fumiyo Fukumoto and Yoshihiro Sekiguchi "Correlating TV news stories with a newswire article" RIAO2000, pp.1372–1380, 2000.
4. Hirotoshi Taira and Masahiko Haruno "Feature Selection in SVM Text Categorization" Trans. of Information Processing Society of Japan,Vol.41,No.4,pp.1113-1123, 2000.
5. E. Brill, "A simple rule-based part of speech tagger", Proceedings of the 3nd conference on applied natural language processing, pp.152–155, 1992.
6. V. Vapnik. "The Nature of Statistical Learning Theory" *Springer, New York*, 1995.
7. J. Allan and J. Carbonell and G. Doddington and J. Yamron and Yang Y., "Topic Detection and Tracking Pilot Study Final Report", the DARPA Broadcast News Transcription and Understanding Workshop, http://www.itl.nist.gov/iaui/894.01/proc/darpa98/ index.htm, 1998.